MOVING FROM DATA AND INFORMATION SYSTEMS TO KNOWLEDGE BUILDING SYSTEMS

ISSUES OF SCALE AND OTHER RESEARCH CHALLENGES

09/30/2003
Version 2.0

NASA

Gail R. McConaughy
Code 586
NASA/GSFC
Greenbelt, MD 20771

Kenneth R. McDonald
Code 586
NASA/GSFC
Greenbelt, MD 20771

ABSTRACT

NASA is acquiring massive volumes of Earth science observations. These remote sensing observations are being processed and transformed into "data" useful for scientific analyses. However, to realize the full potential of these observations the scientific data needs to be further processed and transformed for usage by applications. Handling of such large volumes for the purposes of applying the knowledge will require advances in data systems and analysis techniques. This paper describes how intelligent data understanding algorithms could be an integral component of a next generation of data and information systems; conceived of here as "knowledge building systems" [1], [2]. In order for intelligent algorithms and techniques to play a key role, they will be required to process massive volumes of complex scientific data in a timely fashion. This paper will provide an initial view of the scope and scale of the problem. It is hoped that such a characterization will assist the computer science research community in targeting their research.

TABLE OF CONTENTS

INTRODUCTION

More than ten years ago, NASA embarked on an ambitious program to collect a long-time series of Earth observation data. That program is now well underway, with significant successes. Enormous amounts of data are being returned from a large number of sophisticated sensors. In addition, all this data is being scientifically processed to geophysical parameters and being archived in NASA's active archives in "keep-up" mode. NASA's EOSDIS is doing a good job of processing, storing and providing file level access to this data at a scale and level of service that far exceeds that of earlier NASA missions. However, these volumes are accumulating at a steep rate and are taxing the ability of data archives and users to fully utilize those data. In addition, future flight plans by both NASA and it's partner operational agencies include more advanced instrumentation of even higher spectral and spatial resolution measurements, and a long-term possibility of flying spacecraft and instruments in constellations. For example NOAA and DOD plan a series of spacecraft such as NPOESS and GOES-R that will greatly enhance global coverage. All of this will lead to larger and larger accumulations of long-term data.

NASA's Intelligent Systems/Data Understanding Project[1] is investing in sophisticated algorithms for intelligent understanding. The Intelligent Data Understanding research topic addresses low TRL investigations in the areas of data mining and conditioning, knowledge discovery for scientific understanding and engineering analysis, and machine learning for decision-making and action. These investigations and associated algorithm developments need to be researched in the context of the magnitude of the problem to be solved. This paper will provide an initial snapshot of NASA's current large volumes of scientific data in order to inform the computer science community researching intelligent data algorithms. The next generation of data and information systems are conceived of here as "knowledge building systems". This paper will point to challenges that designers of such systems will face and need to address using scalable and robust intelligent data understanding algorithms

MOVING FROM SCIENCE DATA SYSTEMS TO KNOWLEDGE BUILDING SYSTEMS

Current approaches to science data systems have been, and remain, adequate for some types of scientific research. However, with greater emphasis on the utility of Earth Science data for applications, processing raw collections of data to "actionable" information becomes increasingly important. This requires even higher levels of processing, probably by using intelligent data understanding algorithms to identify events, perform quality assessments [3], and perform both supervised and unsupervised classifications to accumulate content knowledge. Both science research and the application of that science research become drivers of the needs of knowledge building systems. Additionally, if NASA wishes to speed adoption of NASA technology by its partner operational agencies, then NASA should demonstrate the potential for quick utilization of the data it collects. To complete the picture, it is noted that there are parallels in NASA space sciences. For example, some of the deep space missions envisioned, e.g. Prometheus to the moons of Jupiter, will need to process data, identify "features" and act based on what it is sensing.

The software architectures to support this are no longer one-way stovepipe data collection systems. Rather, a new type of system is required – a "knowledge building system". And, "knowledge building systems" will need Intelligent Data Understanding and Intelligent Planning and Scheduling algorithms to be "integral" and "embedded" if they are to deliver on the promise of full data utilization.

This section will provide an overview of current data system types, an overview of envisioned knowledge building systems and a brief description highlighting differences.

AN OVERVIEW OF TODAY'S DATA AND INFORMATION SYSTEMS

The section below discusses the typical set of objects in today's data and information systems. This section also describes a typical interaction between such systems and an end-user employing intelligent data understanding algorithms to support an application.

TODAY'S DATA AND INFORMATION SYTEMS HAVE MADE GREAT STRIDES IN SOLVING THE CHALLENGE OF ACCESS - YET TO BE RESOLVED IS FULL UTILIZATION

Science data systems of today, particularly for Earth science, are already addressing some of the issues of collection, archiving, storing and accessing ever-increasing volumes of data. A "typical" description of the functions of such a system can be seen below.

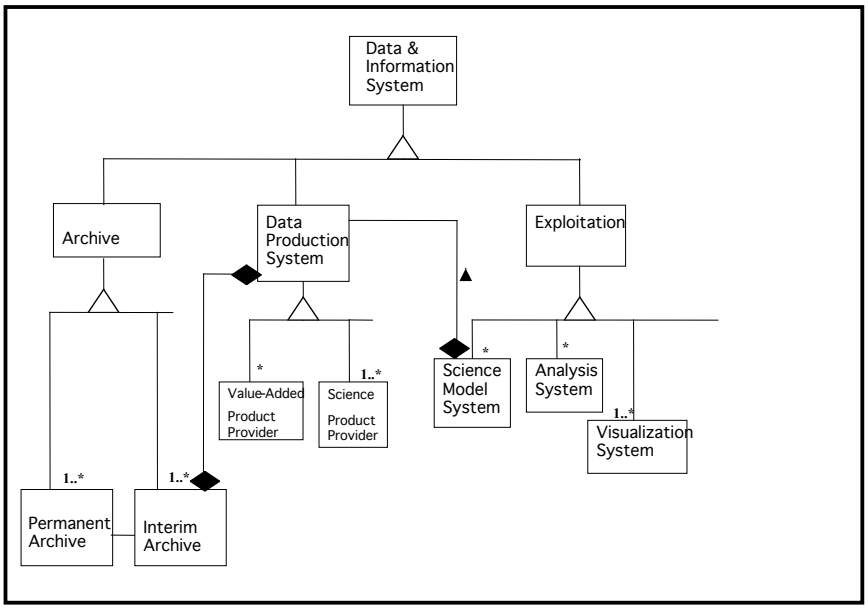


Figure 1: End-to-E End Context of Data and Information Systems

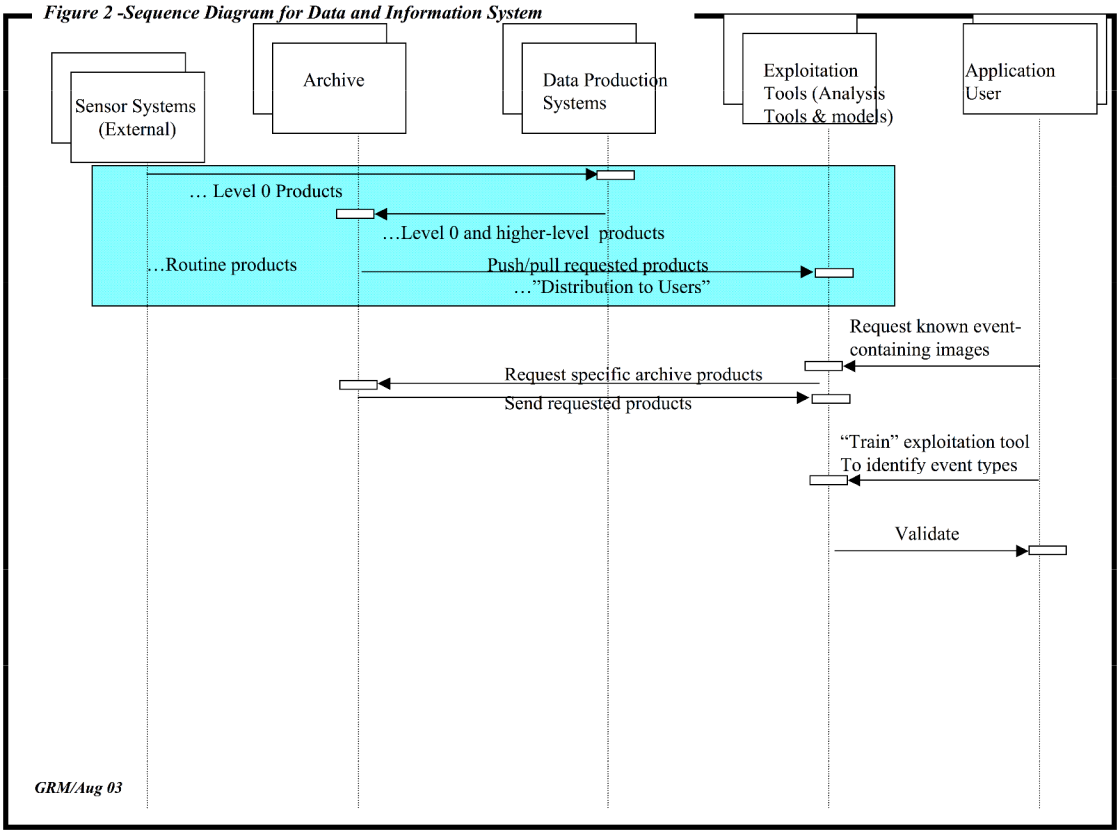
Today's data and information systems, as shown in Figure 1, are primarily composed of archives, both interim and permanent, data production systems that include routine science product production and "value-added" providers, and "exploitation tools" used for visualization, analysis, and modeling of physical processes serving research scientists. The "source" of the data is from a sensor system where often there is a one to one match between the sensor and it's data production system. Archives, however often store more than one collection of data. The reason the "sensor system" is not shown on this diagram is because today's data and information systems are very loosely coupled with the sensor system to the point that it can be considered "external".

The science data production system transforms the data into "corrected" data (that is radiometrically and geometrically corrected products), and then possibly into "science" products that are derived measures of geophysical parameters. These products and the intermediate

products are then archived. While they are actively being quality assured from a scientific viewpoint, they are stored in an "interim" archive. "Active" manipulation of the data by research scientists is via exploitation tools. The data will be visualized and analyzed using data analysis packages, and may possibly be assimilated in to a science modeling system, where such models are mature. After some period the data are then stored in a "permanent" archive if considered necessary for a long-term data record.

The primary emphasis on today's data and information systems is on capture, preservation of the sensor data, and production of derived scientific products to a fixed, pre-planned schedule. In addition today's systems capture pre-planned metadata in order to allow access to those data at some future date by knowledgeable users who were not the original sensor scientists.

"TYPICAL" CONCEPT OF OPERATIONS FOR A DATA AND INFORMATION SYSTEM



A representative concept of operations of today's science data and information systems is shown in Figure 2. Not shown in detail, but requiring the bulk of today's resources are the "routine" and largely "one-way" or "stovepipe" collection, processing and archiving processes leading to distribution of the data to end-users. These are primarily fixed, pre-determined processes. Any iteration or identification of relationships is done external to the data system. During "routine" operations, the focus is on QA, production of geophysical parameter products, preventing loss, and preservation of the very valuable data stream. It is assumed that the scientific researcher is fully capable of dealing with the complexity of the science data. File level access is supplied to scientific researchers via varying services such as search, ftp, subscription, etc.

For the Earth sciences, application users are largely in the experimental phase and are, for the most part, the final process appended far downstream the timeline of data production. Application users typically request importation of routine products from the archive for analysis within exploitation tools resident in their own computing resources, where it is known a priori that a geophysical event has occurred (e.g. fire, algae bloom, cyclone, etc.). A single archive is typically the focus, but searching for data from multiple archives is becoming more common due to the underlying complexity of geophysical events. For example, African dust storms have been linked to algae blooms – iron and nitrogen from African dust storms deposited many hundreds of miles away can provide nourishment to algae. Multiple disciplines would be involved analyzing

such an occurrence. After acquiring the data the application user then interacts with an exploitation tool to "train" the tool to identify statistical signature of known events. Events tend to be examined a scene at a time, with validation done manually by the application user. However, there are a few researchers performing machine-leaning assisted analysis examining limited time series data. After manual interpretation by the application user, alerts may be distributed to the impacted public or policy representatives. For the most part, very large volumes of data are not analyzed automatically.

Today's systems are characterized by the following; one-way stove-pipes, pre-determined scheduling and resource management; experimental usage of intelligent data understanding algorithms at the end of a long string of time-consuming processes, and multiple, redundant "hand-offs" from storage repositories (e.g. satellite, to ground receiving station, to processing archive, active archive, and then finally long-term archive). Overall utilization of the data streams could be characterized as "partial", mainly focused on research scientists analyzing specific phenomena and "experimental" utilization by application users.

AN OVERVIEW OF TOMORROW'S KNOWLEDGE BUILDING SYSTEMS

The section below discusses the envisioned next generation of systems that move from data and information "access" to supporting full "utilization" of as much collected data as possible.

TOMORROW'S KNOWLEDGE BUILDING SYTEMS MUST ADDRESS FULL UTILIZATION IF NASA'S DATA IS TO REALIZE THAT DATA'S UNIQUE VALUE

Given the wide community relevance of NASA's Earth science information Knowledge Building Systems of tomorrow must be able to take the complexity of the science data and make it usable for a myriad of applications. These will include decision support systems with pressing near real time data needs such as flood monitoring and prediction, evacuation route safety assessment, etc.

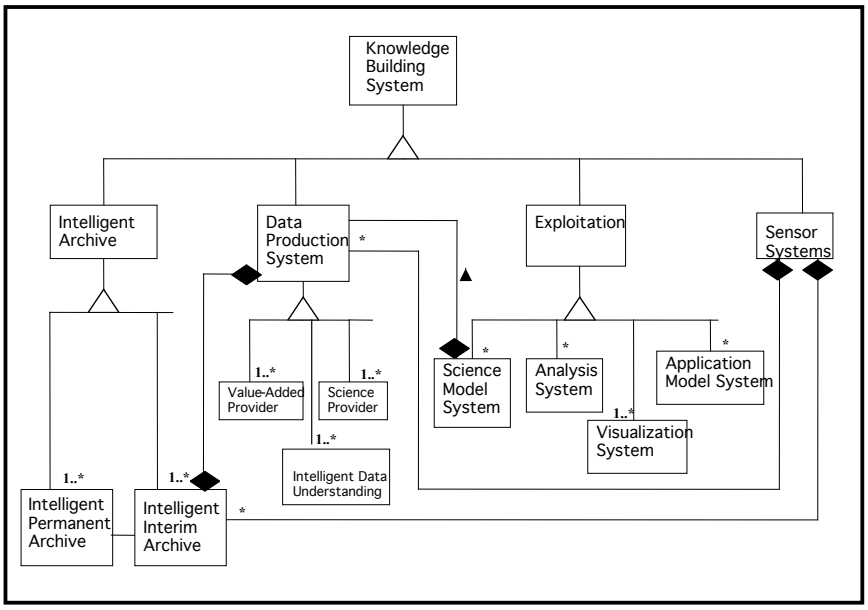


Figure 3: End-to-E End Context of Knowledge Building System

Knowledge Building Systems will differ in some key ways from today's data and information systems. Figure 3 adds "new" objects and functions not found in today's data and information systems. Sensor systems now become "internal" to the system rather than a simple external data source. Sensor systems will be more tightly coupled to the overall system in that its processing and storage may be fully accessible to the rest of the system, and sensor systems may even be considered a "support tool" of the modeling system targeting data needed by the models. Other key differences are that a new type of exploitation tool is added, this being an "application model". An application model would be a model of a specific occurrence of a geophysical event or events, for example a "fire" model. In support of these new objects a knowledge building system also has "intelligent" archives rather than simple storage repositories, and they are likely to be supported by new types of production systems using "embedded intelligent data understanding algorithms". Two key types of infrastructure intelligence, which will surely interact, can be understood to be fundamental. First, "automation" will be a support presence

throughout knowledge building systems that can manage distributed resources with far less manual intervention. Such automation will permit sensor systems to be more flexible in their targeting or mode changes based on science need rather than a fixed pre-determined schedule, etc. Another type of infrastructure "intelligence" of the overall system will be to embed "intelligent data understanding" algorithms within the system, rethinking today's separation of acquisition, processing, storage, followed by analysis. All of these objects need "intelligence" designed in from the start. The following knowledge building system concept of operations will demonstrate how these "embedded intelligent data understanding" algorithms will be executed within a knowledge building system.

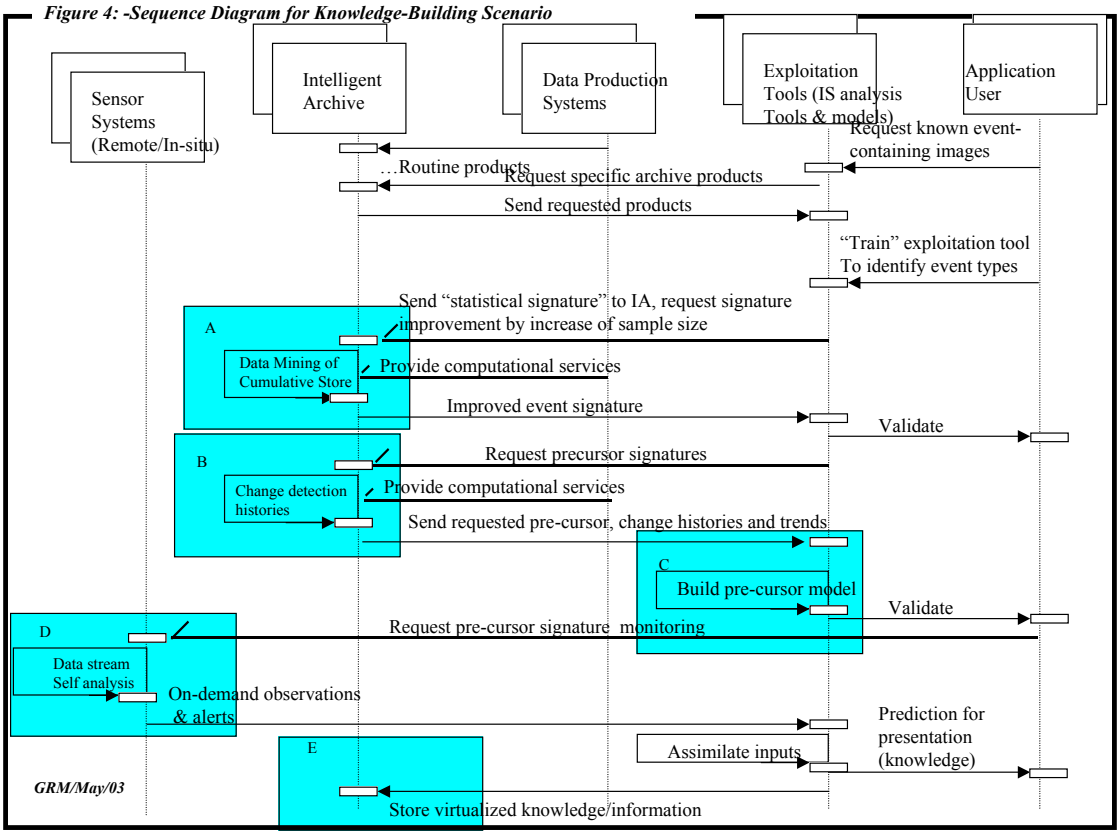


Figure 4 illustrates a Knowledge Building System and highlights new functions and processes.

The "routine" operations of collection, production, archive, and access are the same as today's science data systems and are not shown in detail on this trace diagram. In all likelihood, routine collection, production, archiving and research scientist access will proceed as in today's systems. However, "non-routine" application functions will be permissible and can be accommodated by the more "intelligent" resource management in the new system. To illustrate; as before an application user requests importation of routine products from an Intelligent Archive for analysis within an exploitation tool where it is known a priori that a geophysical event has occurred. That application user interacts with the exploitation tool to "train" the tool to identify the statistical signature of a known event. However, in the new paradigm, the end-user requests that the Intelligent Archive utilize "embedded data understanding" algorithms to help improve that statistical signature by mining large historic archives for similar events (e.g. fires in US South West vs. fires for agricultural clearing…) – "A" in the diagram. The Intelligent Archive builds a database of pointers to remote sensing products containing statistically similar event types. The Intelligent Archive may provide a set of statistical signatures of that event, or may fuse those signatures. In both cases, such a "data base" of statistical signatures based on a large sample set will improve event identification for individual events and multiple events across geographic regions. The application user will likely validate this database. These historic signatures can be

used to support automated event detection as new data is acquired. As a next step ("B"), the Intelligent Archive is requested to "re-mine" the large cumulative archives matching the previously identified statistical signatures and to examine a time series of data to identify relevant pre-cursor signatures. For example, a pre-cursor signature for a fire might be dryness, winds, lightning, etc. For algae blooms it might be previous occurrences of algae blooms, sediment flow, a build up of Sea Surface Temperature, etc. These "pre-cursor" signatures may well have both a temporal and spatial component to understand and record.

These new data mining processes will require "embedded" data understanding – that is, computational services that are provided by data production systems, however, it is very likely that these services need to be "embedded" within the mass storage devices due to their recursive and often computationally intensive natures. By embedded, it is meant that the algorithms likely need to run constantly in the background using non-traditional processing environments. This process of training, mining of historical archives to find like signatures, and then examining those time series to identify pre-cursor signatures will result in the ability to automatically build a "predictive application model" composed of pre-cursor and associated resulting "events" ("C"). As such, in knowledge building systems "application models" become a new addition to the conceptual model with some ability to automatically supply a statistical database of reference information for those models. However, it is important to note that the assumption here is that the "application model" is not the same as a model embodying physics-based computations. Rather, the "application model" consists of databases of statistical patterns that can be used by pattern matching machine learning techniques.

In an era where "sensor-webs" may exist, those pre-cursor signatures may be up-loaded to a collection of sensor systems ("D"), those sensor systems analyze it's incoming observation data stream for pre-cursor event detection, and provide on-demand observations and alerts predicting the possible occurrence of an event. A sensor-web may focus it's resources on the area of interest, collect more and/or higher resolution data, prioritize that data flow to be assimilated into an application model, with a prediction resulting. This means that the sensor systems, along with their local processing and storage, are highly likely to be just a "node" of a knowledge building system. It is anticipated that the same intelligent resource management controlling the archive storage may well manage the sensor systems physical resources. This would reduce the present day redundant, and possibly unnecessary, storage hand-offs making the sensor systems 'integral' to a knowledge building system. To close the loop, all relevant data bases used in mining would be stored to improve future mining requests ("E"). Future deep space missions, may well not rely on accessing long-term archives to mine such data, rather such missions may well be analyzing real-time data streams on acquisition, and automatically reducing that stream to statistical data bases that could be used for system learning and action when response times from controllers introduces too much latency.

Tomorrow's systems will be characterized by the following; network-centric, flexible scheduling and resource management; usage of intelligent data understanding algorithms "embedded" at each key data to knowledge transformation point, and reduced "hand-offs" among storage repositories (e.g. satellite archive may be mirrored in a peer-based ground archive under continuous background processing).

HIGHLIGHTING THE DIFFERENCES

Three key changes between current data and information systems and the envisioned knowledge building systems are critical.

- Greater automation of distributed and heterogeneous data flows across complex resources, moving from "stove pipe" management to "network" management with two-way linkages between collection and prediction.
- Mature, scalable and robust intelligent data understanding algorithms applied in two novel ways – one via near-real time "monitoring" in-coming data streams, and another performing "mining" of decades of cumulative archives.
- New "architectural" concepts that meld collection, archiving, science processing, and automated intelligent data understanding, to reduce "redundant" hand offs, maximizing automated analysis, for example moving from "lazy" data residing idly in archives when it could be being analyzed.

These key advances in capabilities will need to be in place in order to assure that the massive volumes of valuable data collected are utilized to the maximum extent feasible.

CHALLENGES

Each of the identified differences between data and information systems and knowledge building systems will need computer science research if an operational environment is to be developed eventually. This section offers an initial analysis and description of the substantial challenges to be addressed by such research.

VOLUME

Applying intensively recursive 'automated' intelligent data understanding algorithms to the massive volumes currently being accumulated by Earth sciences indicates a need for significant advances in scalability. Two aspects need to be considered: the near real-time transformation of large volumes of raw sensor acquisitions into directly and immediately useful knowledge for an application end-user, and, the mining of a decade of data accumulation.

To get some sense of scale, it is instructive to examine the state of EOSDIS' capabilities for distributing and accumulating very large volumes of data. EOSDIS distribution volumes are indicative of the minimum volumes that future knowledge building systems would process, possibly in near-real time (Figure 4, "D"). Cumulative volumes indicate the scale future knowledge building systems will need to process during data mining (Figure 4 "A" and "B"). The following diagram shows the "routine" production, archiving and distribution volumes in a "snapshot" as of spring 2003.

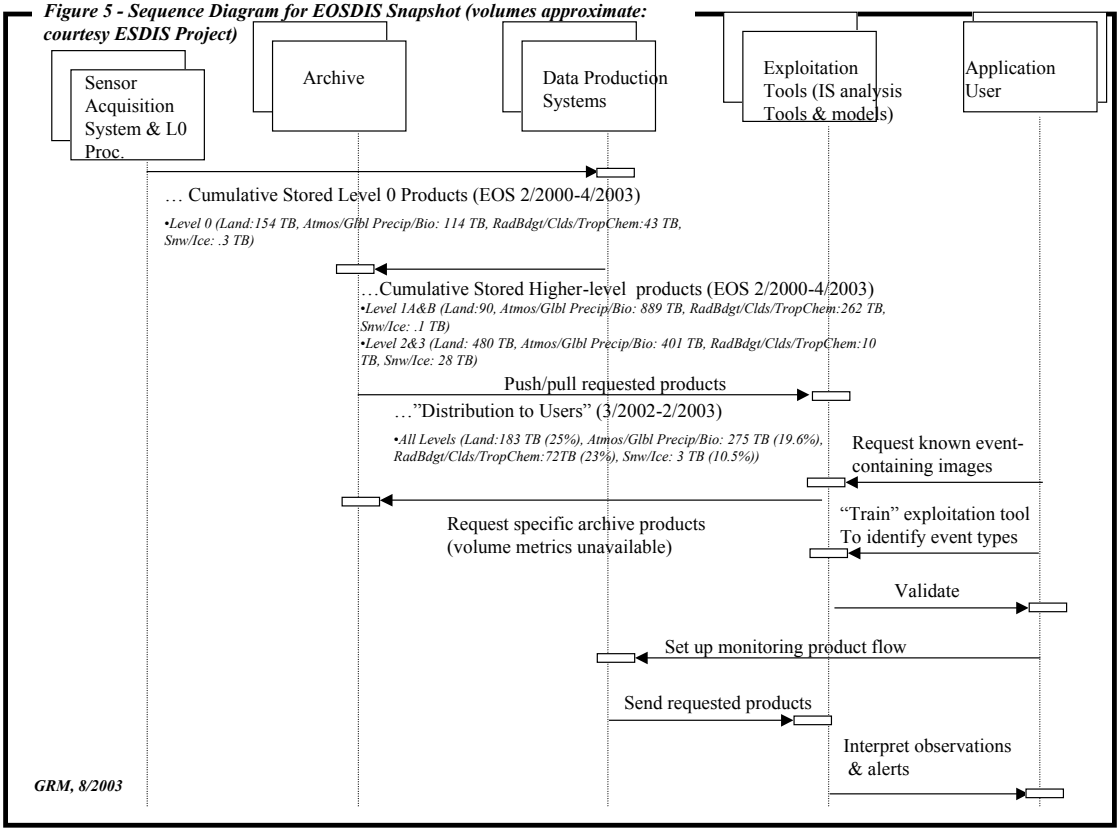


Figure 5 - Sequence Diagram for EOSDIS Snapshot (volumes approximate: courtesy ESDIS Project)

The diagram does not show the entire EOSDIS data holdings or all disciplines, nor does it show "true" data flows. It accounts for some deletion of lower level products after a holding period.

The distribution volumes do not include any metrics for the new EOSDIS function, "data pools". Details of missions and archive centers shown are described in the Appendix.

To summarize, the daily accumulation rates of EOSDIS are on the order of two terabytes per day for the EOS spacecraft. The total accumulation of 3.25 years of EOS data acquisition, beginning with Landsat 7, is approximately 2.5 Petabytes at the end of April 2003 (source: ESDIS Project). The total distribution to users, excluding distribution for the purposes of science processing, test, and QA, for a single year ending in the spring of 2003 was about 533 TeraBytes, or about 22% of the 3.25 year total.

To get a sense of the scale that intelligent data understanding algorithm research should address, assume that tomorrow's Knowledge Building Systems will need to process at least the EOSDIS distribution volumes if they are to translate raw sensor acquisitions into directly and immediately useful knowledge, possibly in near-real time.

Today, in one year of distribution from the 3.25 years of accumulation, EOSDIS distributes 25% Land, 19.6% Atmospheric, Global Precipitation, Biosphere, 23% Radiation Budget, Clouds, Tropospheric Chemistry, and Aerosols, and 10.5% for Snow & Ice. EOSDIS current requirements state that all data products shall be accessible and the system has been sized to distribute 100% of its yearly accumulation. EOSDIS is not sized to distribute data as a function of its multi-year accumulation. However, it is not to be assumed that the solution is simply to increase distribution volumes. Increasing distribution volumes may well not be "throughput" limited (especially as hardware technologies advance over time), it may well be limited by science and application research budgets with too few funded researchers. Rather, the solution is more likely to be decreasing the volume and increasing the knowledge value of what is distributed. That is the purpose behind the envisioned knowledge building system, although the volumes are expected to remain extremely challenging.

For the intelligent data understanding research community to realize some idea of the scope and scale of the technical challenge it may be useful to translate these volumes into units of measure that are more directly interpretable. Since many in the intelligent data understanding community are familiar with Landsat scene volumes the following is provided to give that community some "terms of reference".

- Distribution:
    - If a Landsat ETM + scene is approximately 440MBytes (Land Processes Distributed Active Archive Center (http://edcdaac.usgs.gov/landsat7/l70rwrs.html)), the total distribution to end-users for a single year ending in the spring of 2003 is equivalent to 1.2 Million ETM + scenes.

Given a limited sample, assume a hypothetical case for a highly recursive unsupervised intelligent data understanding algorithm that takes 80 minutes wall-clock of a dedicated machine for one Landsat ETM+ scene. If all of one year's worth of EOS distribution to end-users were to be processed by such an intelligent data understanding algorithm, it will take 184 years.

In previous sections showing knowledge building systems, the idea was introduced that the cumulative archive volumes could be repetitively mined in order to build up a data base of geophysical event statistical signatures (Figure 4, "A" and "B"). Each mining request would have as it's first step accessing those data from the archive. Assume that the cumulative volumes of EOSDIS should be mined in order to understand the volumes that Knowledge Building Systems will need to process to mine historical records.

- Accumulation:

- Three years of data accumulation of Level 1 and higher products is equivalent to 4.9 Million Landsat ETM+ scenes.
- The three-year accumulation of all Level 2 & 3 products is equivalent to 2.1 Million Landsat ETM+ scenes.
- Three years of accumulation of Level 2&3 Land data alone, is equivalent to 1.1 Million Landsat ETM+ scenes.

Needless to say, these dramatic volumes will be an incredible challenge to intelligent data mining of historic data. With the above example, processing all 3 years of data accumulated to date (Level 1 and higher) with such an unsupervised classifier would take over a staggering 700 years, Level 2 & 3 3-year accumulation will take over 300 years. Processing the land data alone, would take just over 160 years. If the hypothetical example is representative of many intelligent data understanding algorithms, the processes in Figure 4, "A" and "B" that require mining historical stores would be impossible.

EOSDIS will continue accumulating data through at least 2010 at approximately 1 PetaByte/year (the lifetime of the Aura mission if launched Jan 2004). These extremely large volumes coupled with the large processing times indicated in the above paragraph point to a serious need to consider scale when researching algorithms and techniques. As volumes accumulate, much greater disparities will be seen between distribution and cumulative storage. Research challenges are numerous. How will the cumulative archive be "processable" to add to a reanalyzed historic record? How will valuable content be identified if it is not a priori known to exist? Will content be identified by examining a series of Level 3 data? What if phenomena are lost in the reduction of resolution between Level 2 and 3? How can intelligent data understanding algorithms be applied to help address these issues? Could content statistics be collected as the Level 2 is being archived? Could it be collected via background processes, rather than allow the data to be "lazy" by just sitting unused in an archive? Could new concepts need to be developed to "embed" such recursive algorithms in every possible site to permit them to run constantly, in the background in a highly parallel fashion? It should always be in a state of being analyzed, if not scientifically, then at least statistically.

Ultimately, in order to support full utilization of the data, the question becomes one that indicates that archives should not simply be "access" engines, they need to be actively transforming the data to knowledge, and hopefully reducing the volumes that are distributed while increasing the knowledge value of the information that is distributed.

TIMELINESS

Current production to archive rates are proceeding at "keep up rates", which is a substantial achievement over prior programs. Still, given all the redundant hand-offs from data store to data store, the true time line starting from acquisition to archive takes on the order of "days", with science-quality QA taking on the order of "months" to "years". Some decision support applications have timeliness requirements on the order of "minutes" from acquisition to event identification and location. Daily accumulation rates of all new products at all levels are currently proceeding at about a total of about 1 PetaByte per year, averaging 2.8 TeraBytes per day. This is equivalent to about 2.3 Million ETM+ scenes a year, and 6.4 Thousand a day. Combining applications specific processing via an intelligent-data-understanding algorithm with substantially shorter time requirements poses another daunting issue of "scale" for the intelligent data understanding algorithm developer. Using the above 80 minutes per IDU analysis, "keep up rates" would require compressing a processing requirement of just under a year of processing time into a single day. Again, without advances the processes described in Figure 5 "D" become impossible.

FULL UTILIZATION

HIGHLY PROCESSED DATA IS REQUIRED TO ENGAGE MORE APPLICATION USERS

Current application usage of NASA's data stream could be characterized currently as largely "applied research". A number of very promising early results point to ever broadening applicability (for example, identifying fires in MODIS data (http://activefiremaps.fs.fed.us/index.html)). Could intelligent data understanding techniques be used to process complex scientific data into forms that would be more directly usable to the applications community?

NEW PARADIGMS FOR LONG TERM MINING

We are now actively building a massive long-term record of many key geophysical parameters, with but a trickle of data going to experimental application users. Should we be waiting for a scientist to get funding to do some future "pathfinder" exercise to do a massive reprocessing campaign to look for a phenomena of interest, or should we be constantly mining the data statistically? Statistical signatures could be mined, and stored, and then later used within an analysis process. Would this provide greater utility of the data on more possible topics? Should the data be mined statistically, as well as processed scientifically to make the long-term accumulation more fully utilized?

NEW PARADIGMS FOR AUTOMATED "APPLICATION MODEL" BUILDING

Constant background search for precursor signatures of significant events may well help to build a long-term statistical data base that could be an invaluable resource, and may well lead to new ways of building a "model" that could be used in a predictive capacity (Figure 5 "C"). Could databases of "pre-cursor" signatures be used in a 'predictive' sense, again increasing the value of these massive data stores?

CONCLUSION

The intent of this paper is to analyze the potential "place" that intelligent data understanding algorithms could have in a very advanced next generation "knowledge building system". However, there remain substantial challenges with scaling such algorithms for the real operational environment that NASA is facing. As such, this paper attempts to acquaint the potential computer science researcher with the issues of scale associated with the large volumes of EOS data currently being, and planned to be, scientifically processed and stored within EOSDIS.

Further research topics that are planned to be pursued include; surveying performance issues across multiple types of machine learning techniques, analyzing current experimental uses of intelligent data understanding algorithms for how they might decrease data distribution while increasing utilization, and finally examining in more detail the role that intelligent data understanding algorithms might play in automated application model building.

Reference:

[1] Ramapriyan, H.K., G. McConaughy, C. Lynnes, S. Kempler, K. McDonald, R. Harberts, L. Roelofs, and P. Baker, 2002. "Conceptual Study of Intelligent Archives of the Future", Report prepared for the Intelligent data Understanding program, 39 p., http://daac/IDA/IA_report_8-27-02_baseline.pdf.

{2} H. K. Ramapriyan, et al, "Intelligent Archive Concepts for the Future", *Proceedings of the ISPRS/Future Intelligent Earth Observing Systems Conference*, Denver, CO, November 2002.

[3] Isaac, D. and Christopher Lynnes, 2003. *Automated Data Quality Assessment in the Intelligent Archive*, White Paper prepared for the Intelligent Data Understanding program, 17 p.,http://daac/IDA/http://daac.gsfc.nasa.gov/IDA/presentations.shtml

APPENDIX

Table 1 Data Level Definitions

Level 0	Instrument data at original resolution, time order-restored, with duplicate packets removed
Level 1A	Level 0 data that are reformatted with calibration data and other ancillary data included. Geolocation information for each spatial element (e.g., pixel) of the reformatted sensor-coordinate data is stored separately.
Level 1B	Level 1A data to which the radiometric calibration algorithms have been applied to produce radiances, irradiances, or brightness temperature.
Level 2	Geophysical parameter data retrieved from a single sensor's Level 1B data by application of geophysical parameter algorithms. The MODIS science team has an additional level, 2G, which contains pixel to grid mappings.
Level 3	Earth-gridded geophysical parameters that have been averaged, gridded, or otherwise rectified or composited in space and/or time.
Level 4	Model output or results of analyses from lower-level data; such as variables derived from data collected by multiple sensors

Short guide to reading UML diagrams

Definitions:

Class diagram: describes the static structure of a system, how it is structured rather than how it behaves. Contain following elements:

- Classes: represent entities with common characteristics or features; features include attributes, operations and associations.
- Associations: represent relationships that relate two or more other classes where the relationships have common characteristics or features. These features include attributes and operations. Some types of associations:
  -- Line ending with filled diamond: compositions relationship (contains, if you remove one you must remove the other)
  -- Line ending with hollow diamond: aggregation relationship (removal of one does not mean removal of another – can exist standalone)
  -- Numbers on lines: 1 (one involved in relationship), 1…* (one or more involved in relationship), * (zero or more)
  --Arrow: the direction in which to read the association name (otherwise read from top to bottom, from left to right)
  -- Hollow arrow (looks like a triangle): inheritance (lower items inherit or receive all the characteristics of upper item), generalization, "is-a-kind-of"

Object Diagrams: describe static structure of a system at a particular time, class diagram describes all possible situations, an object model describes a particular situation. Contains the following elements:

- Objects: particular entities. These are instances of classes.

- Links: particular relationships between objects, instances of associations.

Example of a class diagram (object not shown)

Reading Sequence Diagrams:

Sequence Diagrams: Describe interactions among classes. Modeled as exchanges of messages. Focus on classes and messages they exchange to accomplish some desired behavior. A particular instance of a sequence diagram is called a scenario. Used to elaborate use cases. Elements of a sequence diagram:

- Class roles: roles that objects may play within the interaction (names of objects at the top of the diagram)
- Lifelines: represent the existence of an object over a period of time (vertical lines extending down from each object)
- Activations: represent the time during which an object is performing an operation (thin rectangles on lifelines). Ended in full arrow shows synchronous messages, half arrows are asynchronous messages.
- Message: represent communication between objects (horizontal arrows – labeled with the message being sent between the class roles, triggers an operation in the receiving object)

Figure 5, Page 15, Detailed Description

Flight missions that are shown: Landsat 7 (Launch Apr-99), Terra (Dec-99), Meteor3/Sage (Dec-01), ACRIM (Dec-99), Aqua (May-02), Midori-II/AMSR (Dec-02) (to be launched in 03, ICESat/GLAS (Jan-03), SORCE (Jan-03), Aura (Jan-04)). Flows are broken by DAAC, with an assumed set of disciplines. Where a DAAC has an aggregation of disciplines, the cumulative archive volumes are not broken down: EDC (Land), GSFC (Atmosphere, Global Precipitation, Biosphere), NSIDC (Snow & Ice), LaRC (Radiation Budget, Clouds, Tropospheric Chemistry, Aerosols)

“Distribution to Users” is the total product volume delivered to end users via orders or subscriptions for a year ending in the spring of 2003 (total to end users 533TB), this excludes QA, test, and production. (source: ESDIS Project). These figures do not include distribution from a new function in the process of being added to EOSDIS, the “data pools”.